Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique

Didier Bourigault, Cécile Frérot

Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique

Didier Bourigault et Cécile Frérot

ERSS – CNRS & Université Toulouse le Mirail
5, allées Antonio Machado
31 058 Toulouse Cedex 1
{didier.bourigault,cecile.frerot}@univ-tlse2.fr

Mots-clés : analyse syntaxique, ambiguïté de rattachement prépositionnel, sous-catégorisation syntaxique
Keywords: syntactic parsing, PP attachment disambiguation, subcategorization lexicon

Résumé

Cette étude est menée dans le cadre du développement de l'analyseur syntaxique de corpus Syntex et porte sur la tâche de désambiguïsation des rattachements prépositionnels. Les données de sous-catégorisation syntaxique exploitées par Syntex pour la désambiguïsation se présentent sous la forme de probabilités de sous-catégorisation (que telle unité lexicale - verbe, nom ou adjectif - se construise avec telle préposition). Elles sont acquises automatiquement à partir d'un corpus de 200 millions de mots, étiqueté et partiellement analysé syntaxiquement. Pour évaluer ces données, nous utilisons 4 corpus de test de genres variés, sur lesquels nous avons annoté à la main plusieurs centaines de cas de rattachement prépositionnels ambigus. Nous testons plusieurs stratégies de désambiguïsation, une stratégie de base, une stratégie *endogène* qui exploite des propriétés de sous-catégorisation spécifiques acquises à partir du corpus en cours de traitement, une stratégie *exogène* qui exploite des propriétés de sous-catégorisation génériques acquises à partir du corpus de 200 millions de mots, et enfin une stratégie *mixte* qui utilisent les deux types de ressources. L'analyse des résultats montre que la stratégie mixte est la meilleure, et que les performances de l'analyseur sur la tâche de désambiguïsation des rattachements prépositionnels varient selon les corpus de 79.4 % à 87.2 %.

Abstract

We carry out an experiment aimed at using subcategorization information into a syntactic parser for PP attachment disambiguation. The subcategorization lexicon consists of probabilities between a word (verb, noun, adjective) and a preposition. The lexicon is acquired automatically from a 200 million word corpus, that is partially tagged and parsed. In order to assess the lexicon, we use 4 different corpora in terms of genre and domain. We

assess various methods for PP attachment disambiguation : an exogeous method relies on the sub-categorization lexicon whereas an endogenous method relies on the corpus specific ressource only and an hybrid method makes use of both. The hybrid method proves to be the best and the results vary from 79.4 % to 87.2 %.

1 Introduction

Les nombreux travaux sur le développement de parseurs statistiques concernent la langue anglaise et tendent à utiliser comme corpus d'apprentissage et comme corpus de test des portions de la section du Wall Street Journal du Penn TreeBank (Charniak, 1997). Outre qu'elle permet d'éviter la tâche laborieuse de construction de corpus annotés, cette démarche présente l'immense avantage de pouvoir comparer les parseurs entre eux (Ratnaparkhi *et al.*, 1994 ; Pantel et Lin, 1998). Cette exploitation mono-corpus pose cependant la question de la stabilité des performances en fonction du type de corpus, comme le mentionnent Kilgarrif et Grefenstette (2003 :341) : « *there is little work on assessing how well one language model fares when applied to a text type that is different from that of the training corpus* ». Par ailleurs, il est maintenant bien connu que, dans tout corpus, certaines unités lexicales ont des propriétés syntaxiques de sous-catégorisation spécifiques, qui peuvent donc varier d'un domaine à l'autre (Roland, Jurafsky, 1998 ; Basili *et al.*, 1999). Or peu de travaux relatent des expériences sur la variation des performances de l'analyseur en fonction du type de corpus à traiter, sur le problème de la possible variation inter-corpus et sur celui de la nécessaire adaptation des règles de l'analyseur à un corpus donné. On peut néanmoins citer (Sekine, 1997 ; Gildea, 2001 ; Slocum, 1986).

Dans cet article, nous nous intéressons à l'acquisition et à l'évaluation sur corpus de données de sous-catégorisation syntaxique. Cette étude est menée dans le cadre du développement de l'analyseur syntaxique de corpus Syntex et porte sur la tâche de désambiguïsation des rattachements prépositionnels[1] (section 2). Les données de sous-catégorisation syntaxique exploitées par Syntex pour la désambiguïsation se présentent sous la forme de probabilités de sous-catégorisation (que telle unité lexicale - verbe, nom ou adjectif – se construise avec telle préposition). Dans la section 3, nous décrivons comment elles sont acquises automatiquement à partir d'un corpus de 200 millions de mots, étiqueté et partiellement analysé syntaxiquement. La section 4 est consacrée à l'évaluation sur des données acquises sur 4 corpus de test de genres variés, sur lesquels nous avons annoté à la main plusieurs centaines de cas de rattachement prépositionnels ambigus. Dans la section 5, nous présentons plusieurs stratégies de désambiguïsation : une stratégie de base, une stratégie *endogène* qui exploite des propriétés de sous-catégorisation spécifiques, acquises à partir du corpus en cours de traitement, une stratégie *exogène* qui exploite des propriétés de sous-catégorisation génériques, acquises à partir du corpus de 200 millions de mots, et enfin une stratégie *mixte* qui utilise les deux types de ressources. L'analyse des résultats (section 6) montre que la stratégie mixte est la plus performante, et que les performances de l'analyseur sur la tâche de désambiguïsation des rattachements prépositionnels varient selon les corpus.

---

[1] Nous nous intéressons dans cet article aux prépositions autres que *de*. Le traitement de la préposition *de* repose sur les mêmes principes, mais est sensiblement plus complexe (Frérot *et al.*, 2003).

2 Syntex, un analyseur syntaxique de corpus

Cette expérience est menée dans le cadre du développement de l'analyseur syntaxique de corpus Syntex (Bourigault, Fabre, 2000). Syntex est un analyseur en dépendance qui prend en entrée un corpus de phrases étiquetées[2], et calcule pour chaque phrase les relations de dépendance syntaxique entre les mots. C'est un analyseur en couches (Aït-Moktar et al., 2002) : le corpus est analysé en plusieurs passes, différents modules prenant successivement en charge une relation syntaxique de dépendance donnée, et les sorties d'un module constituant les entrées du module suivant. Chaque module est constitué d'un ensemble de règles construites « à la main ».

Syntex est un analyseur semi-lexicalisé. Le module qui effectue les rattachements prépositionnels exploite des données lexico-syntaxiques de sous-catégorisation, exprimées sous la forme de probabilités qu'une unité lexicale donnée (verbe, nom, adjectif) se construise avec telle ou telle préposition. Le rattachement des prépositions à leur recteur s'effectue en deux passes : (1) recherche des candidats recteurs, (2) choix d'un recteur. Un premier module (*rechercher-candidats*) traite l'ensemble des phrases du corpus, et recherche pour chaque préposition, le ou les mots susceptibles de régir cette préposition. Ce module est constitué de règles qui reconnaissent un certain nombre de configurations linéaires de mots et de catégories morphosyntaxiques à gauche de la préposition au sein desquelles sont identifiés des mots susceptibles de régir la préposition. Ces règles s'appuient sur les relations de dépendance placées par les modules antérieurs, et sont capables d'aller chercher des candidats recteurs dans des configurations relativement complexes, incluant par exemple des structures coordonnées ou des incises. Les configurations d'ambiguïtés, définies comme la succession des catégories grammaticales des candidats recteurs, sont très variées. Sur les 4 corpus de test présentés dans la section 4, la configuration 'V N', où seuls un verbe et un nom sont en compétition - configuration traitée dans beaucoup de travaux dont ceux, fondateurs, de Hindle et Rooth (1993) -, ne représente que 50 % des cas dans le corpus littéraire, environ 35 % dans le corpus journalistique et 15 % dans le corpus juridique et le corpus technique.

Au cours de la seconde étape du traitement des ambiguïtés prépositionnelles, le second module (*choisir-candidat*) revient sur chaque cas ambigu et choisit le recteur de la préposition parmi les candidats. Pour ce faire, ce module exploite des informations de sous-catégorisation associées aux couples (candidat, préposition). Depuis l'origine de nos travaux sur l'analyse syntaxique, ces informations sont acquises de façon endogène sur le corpus en cours de traitement (Bourigault, 1993). En effet, l'analyseur est utilisé dans différents contextes applicatifs, et principalement dans des applications de construction de terminologies ou d'ontologies spécialisées à partir de textes. Il traite des corpus spécialisés, thématiques, de taille moyenne (quelques centaines de milliers de mots, sur des domaines techniques, juridiques, médicaux). Les expériences menées sur de nombreux corpus ont montré que ces corpus renferment des spécificités lexicales, en particulier que certains mots, fréquents dans le corpus, manifestent des comportements syntaxiques spécifiques et imprédictibles. C'est pourquoi, nous avons porté nos efforts depuis une dizaine d'années sur le développement de procédures d'apprentissage endogène sur corpus qui permettent à l'analyseur d'acquérir lui -

[2] Nous utilisons actuellement les versions française et anglaise du Treetager (http://www.ims.uni-stuttgart.de)

même, par analyse du corpus à traiter, des informations de sous-catégorisation spécifiques à ce corpus, acquises à partir des cas non ambigus repérés par le module *rechercher-candidats*.

Devant les limites inhérentes à l'exploitation d'informations de sous-catégorisation acquises exclusivement sur le corpus en cours de traitement, nous travaillons à l'élaboration de ressources générales, susceptibles d'être exploitées pour tout corpus (Frérot *et al*., 2003). Nous avons expérimenté l'utilisation d'un lexique de sous-catégorisation construit à partir des tables du *Lexique Grammaire* (Frérot, à paraître). Nous présentons dans ce travail une expérience d'acquisition de probabilités de sous-catégorisation à partir d'un corpus de 200 millions de mots.

3 Acquisition de propriétés de sous-catégorisation à partir d'un corpus de 200 millions de mots

Les méthodes d'acquisition de propriétés de sous-catégorisation exploitent classiquement des corpus étiquetés de grande taille (Basili, Vindigni, 1998). Le Web est aussi considéré comme source potentielle d'acquisition (Gala Pavia, 2003). Dans notre étude, nous utilisons comme base d'apprentissage un corpus de 200 millions de mots, constitué des articles du journal *Le Monde*, des années 1991 à 2000 (corpus LM10[3]). Nous ne prétendons pas que ce corpus soit représentatif de la « langue générale », mais nous considérons que sa taille et sa diversité thématique en font un corpus référentiellement et linguistiquement peu marqué, à partir duquel il est possible d'acquérir des données de sous-catégorisation relativement génériques. La procédure d'acquisition est adaptée des méthodes d'apprentissage endogène intégrées dans Syntex. La méthode de calcul des probabilités de sous-catégorisation s'appuie sur un ensemble de triplets (recteur, préposition, régi) extraits d'une analyse syntaxique du corpus LM10 effectuée par Syntex. La procédure d'acquisition se déroule en deux étapes, au cours desquelles la même méthode de calcul de probabilités est lancée successivement sur deux ensembles différents de triplets : une étape d'amorçage et une étape de consolidation.

Au cours de l'étape d'amorçage, le module *rechercher-candidats* traite l'ensemble du corpus LM10, qui a été analysé par les modules antérieurs de Syntex, et construit, à partir des cas non-ambigus, c'est-à-dire ceux pour lesquels il n'a identifié qu'un seul candidat recteur pour la préposition, un ensemble de triplets (w,p,w'), où w est le recteur de la préposition p, et w' le mot (nom, ou verbe à l'infinitif) régi par la préposition. Le module *rechercher-candidats* compte aussi pour chaque mot w le nombre d'occurrences dans le corpus où ce mot n'est candidat *d'aucune* préposition. A l'issue du traitement de l'ensemble du corpus, on dispose des données de fréquence suivantes :

- F(w,0) : nombre d'occurrences non ambiguës où le mot w ne régit aucune préposition ;

[3] Ce corpus a été préparé, à partir de fichiers obtenus auprès de l'agence Elra, par Benoît Habert (LIMSI), qui a effectué les tâches de nettoyage, de balisage et de signalisation nécessaires pour transformer les fichiers initiaux en un corpus effectivement « traitable » par des outils de Traitement Automatique des Langues. Nous remercions Benoît Habert et le LIMSI de nous avoir permis de bénéficier de cette ressource.

- F(*w*,*p*,*w'*) : nombre d'occurrences non ambiguës où le mot *w* régit la préposition *p*, qui elle-même régit le mot *w'*.

A partir de ces données, un premier ensemble de probabilités de sous-catégorisation P(w,p) est calculé, selon la méthode décrite plus loin dans la présente section.

Au cours de l'étape de consolidation, le module *choisir-candidat* exploite ce premier lexique et traite à son tour l'ensemble du corpus LM10, analysé par le module *rechercher-candidats*. Il revient sur les cas ambigus et choisit le candidat recteur dont la probabilité de construction avec la préposition, fournie dans le premier lexique, est la plus importante. A partir de ces nouvelles annotations, un nouvel ensemble de triplets est constitué, qui inclut le précédent et auquel s'ajoutent les triplets (*w*,*p*,*w'*) issus des cas ambigus résolus. De nouvelles données de fréquence F(*w*,*p*,*w'*) et F(*w*,0) sont alors constituées, à partir desquelles un second ensemble de probabilités de sous-catégorisation est calculé, selon la méthode décrite ci-dessous. C'est le lexique construit à l'issue de cette étape de consolidation qui est utilisé dans Syntex.

La méthode de calcul des probabilités est simple. La probabilité est calculée comme une fréquence relative pondérée[4]. Soit T, l'ensemble des triplets (*w*,*p*,*w'*), obtenu à l'issue de l'étape d'amorçage ou à celle de consolidation. Pour un couple (*w*,*p*), on définit $E_{w,p}$ comme l'ensemble des mots *w'* tels que la fréquence F(*w*,*p*,*w'*) est supérieure à 0. On définit la *productivité* du couple (*w*,*p*), Prod(*w*,*p*), comme le cardinal de l'ensemble $E_{w,p}$, c'est-à-dire comme le nombre de mots *différents* que régit la préposition *p* quand elle-même est régie par le mot *w*. Nous utilisons ce coefficient pour pondérer la fréquence totale du couple (*w*,*p*). A fréquence égale, plus le couple (*w*,*p*) a été repéré avec des contextes *w'* différents, plus grande est estimée la propension du mot *w* à régir la préposition *p*. L'expérience montre que, dans des corpus thématiques, la très haute fréquence de certains syntagmes très répétitifs incluant le triplet (*w*,*p*,*w'*) vient biaiser la probabilité d'association lexicale entre *w* et *p*. La pondération proposée ci-dessus vise à limiter une telle surestimation et à accorder un poids non seulement à la fréquence de l'association, mais aussi à sa diversité. La formule de calcul de la probabilité pondérée est donnée dans le tableau 1 : F(*w*,*p*) est la fréquence totale du couple (*w*,*p*), F(*w*) est la fréquence totale du mot *w*, et • est un coefficient de normalisation, choisi de telle sorte que la somme des probabilités associées à un mot donné soit égale à 1.

[4] Nous n'avons pour le moment pas testé d'autres méthode de filtrage, comme celle de la distribution polynomiale (Manning, 1993).

T = { (w,p,w') / F(*w*,*p*,*w'*) > 0 }, ensemble de triplets
F(*w*,*p*,*w'*) : nombre de cas où le mot *w* régit la préposition *p*, elle-même régissant le mot *w'*
F(w,0) : nombre de cas où le mot *w* ne régit aucune préposition
$E_{w,p}$ = { w' / F(*w*,*p*,*w'*) > 0 }, le contexte du couple (*w*,*p*)
Prod(*w*,*p*) = Card($E_{w,p}$), la productivité du couple (*w*,*p*)
F(*w*,*p*) = $\bullet_{w' \bullet Ew,p}$ F(*w*,*p*,*w'*)
F(*w*)=F(*w*,0)+ $\bullet_{p}$ F(*w*,*p*)
P(*w*,0) = F(*w*,0)/F(*w*)
P(*w*,*p*)=F(*w*,*p*) /F(w)*log(1 + Prod(*w*,*p*))/ •

Tableau 1. Méthode de calcul des probabilités de sous-catégorisation

Le nombre total d'occurrences de triplets (*w*,*p*,*w'*) à partir desquels les probabilités sont calculées est de l'ordre de 6,7 millions à l'issue de l'étape d'amorçage, et de 12 millions à l'issue de l'étape de consolidation. Le nombre total d'occurrences de mots ne régissant pas de préposition est d'environ 87 millions à l'issue de l'étape d'amorçage , et de 95 millions à l'issue de l'étape de consolidation. Les probabilités ne sont calculées que pour les couples (*w*,*p*) tels que la fréquence totale du mot *w* est supérieure à 20. Un couple n'est retenu dans le lexique de désambiguïsation que si la probabilité dépasse le seuil de 0.01. Le lexique final compte 6 693 verbes différents (chacun pouvant être présent avec plusieurs prépositions), 11 528 noms et 698 adjectifs.

4 Annotation

De façon générale, le développement d'un analyseur syntaxique robuste exige une méthode de travail qui assume la très grande variabilité des corpus sur le plan syntaxique. Les stratégies et règles des différents modules de Syntex sont à chaque expérimentation élaborées à partir de tests effectués sur plusieurs corpus, aussi diversifiés que possible, pour limiter les biais d'implémentation que pourrait introduire une approche mono-corpus. A la variabilité inter-corpus, il faut ajouter la variabilité intra-corpus. Pour éviter d'élaborer des règles trop dépendantes de telle ou telle configuration syntaxique ou unité lexicale, il faut sur chaque corpus annoter à la main un très grand nombre de cas. Dans le cadre de cette étude, nous avons évalué le lexique de sous-catégorisation sur 4 corpus de test, de genres variés, dans lesquels nous avons validé à la main plusieurs centaines de cas :

- BAL. Le roman « Splendeurs et misères des courtisanes », d'Honoré de Balzac (199 789 mots) : 672 cas validés
- LMO. Un extrait du journal *Le Monde* (673 187 mots) : 1 238 cas validés

- REA. Un corpus de comptes-rendus d'hospitalisation dans le domaine de la réanimation chirurgicale (377 967 mots) : 646 cas validés

- TRA. Le *Code du travail* de la législation française (509 124 mots) : 1 150 cas validés

Les règles d'annotation sont les suivantes : (1) ne pas valider de cas où il y a des erreurs d'analyse des modules antérieurs, en particulier des erreurs d'étiquetage, autrement dit on évalue le module de rattachement prépositionnel dans des contextes où les informations sur lesquelles il s'appuie sont justes ; (2) se donner la possibilité de retenir comme valides deux recteurs pour une préposition donnée, en particulier pour les constructions à verbe support (*apporter une aide à*) ; (3) ne pas valider certains cas trop répétitifs, afin de ne pas sur représenter un cas trop spécifique au corpus, comme par exemple dans le corpus CTRA, où les cas de rattachement des participes passés à la préposition sont massifs (ex: *définir les modalités visées à l'article*) ; (4) valider de manière indifférenciée des groupes prépositionnels arguments ou circonstants. Ce dernier point est important, et peut prêter à controverse, si on ne replace pas la tâche d'annotation dans le contexte de l'évaluation des performances d'un analyseur syntaxique. La distinction argument/circonstant, ou complément essentiel/complément circonstanciel, ne fait pas l'objet d'un consensus dans la communauté linguistique. En dehors des cas trivaux, choisis en général soigneusement pour illustrer cette distinction, la confrontation avec des énoncés réels met à mal la clarté de cette distinction (Fabre, Frérot, 2002). Dans ces conditions, la tâche essentielle dévolue à l'analyseur est d'abord de choisir le bon recteur parmi un ensemble de recteurs possibles, et ensuite seulement, et éventuellement, de distinguer le type de complément.

5 Méthode de désambiguïsation

L'algorithme de désambiguïsation mis en œuvre dans le module *choisir-candidat* est simple. Nous comparons 4 stratégies différentes, selon le type des données de sous-catégorisation qu'elles exploitent.

- Mode *base*. En mode base, le module *choisir-candidat* se contente de choisir comme recteur le premier candidat dans l'ordre linéaire de phrase, c'est-à-dire le plus éloigné de la préposition[5].

- Mode *exogène*. En mode exogène, le module *choisir-candidat* exploite le lexique de sous-catégorisation construit à partir du corpus LM10 (section 3). Il choisit le candidat dont la probabilité est la plus élevée. On distingue exogène 1 et exogène 2, selon que le lexique utilisé est obtenu après la phase d'amorçage ou après la phase de consolidation.

- Mode *endogène*. En mode endogène, le module *choisir-candidat* exploite le lexique de sous-catégorisation construit à partir du corpus en cours d'analyse[6]. Avant

[5] Globalement - sur l'ensemble des corpus et sur l'ensemble des configurations d'ambiguïté -, cette stratégie est meilleure que celle qui choisirait le candidat le plus proche.

[6] Selon la méthode décrite dans la section 3, sans l'étape de consolidation.

d'exploiter les probabilités de sous-catégorisation, il exploite la liste des fréquences des triplets (*w*,*p*,*w'*) construite par le module *rechercher-candidats* : si *p* est la préposition et *w'* le mot qu'elle régit, le module choisit le candidat w_i pour lequel la fréquence F(w_i,*p*,*w'*) est la plus élevée. Sinon, il choisit le candidat dont la probabilité endogène est la plus élevée.

- Mode *mixte*. Le mode mixte est analogue au mode endogène, à ceci près que le module *choisir-candidat* choisit le candidat qui a la probabilité endogène *ou* la probabilité exogène la plus élevée.

Dans tous ces modes, la règle par défaut est celle de la stratégie de base, à savoir le choix du premier candidat.

	BAL	LMO	REA	TRA
base	83.0	70.3	59.9	65.5
endogène	83.5	80.1	78.0	82.3
exogène 1	85.7	85.5	65.3	85.9
exogène 2	86.9	86.6	66.3	86.3
mixte	86.6	85.9	78.3	87.3

Tableau 2. Taux de précision (%) des différentes stratégies de désambiguïsation sur les 4 corpus de test

6 Résultats et discussion

Le tableau 2 donne les taux de précision des différentes stratégies de désambiguïsation sur les 4 corpus de test. On peut rapprocher ces résultats de ceux, récapitulés dans (Pantel et Lin, 1998), obtenus sur 3 000 cas ambigus extraits de la partie Wall Street Journal du Penn TreeBank par différentes méthodes : 81,6% avec une méthode supervisée utilisant un modèle d'entropie maximale (Ratnaparkhi *et al*., 1994), 88,1% avec une méthode supervisée utilisant un dictionnaire sémantique (Stetina, Nagao, 1997) et 84.3% avec une méthode non supervisée utilisant des mots distributionnellement proches (Pantel, Lin, *op.cit*.). Etant donné que les langues, le type de corpus de test et les conventions d'annotations sont différentes, il est délicat de comparer ces chiffres avec ceux que nous présentons dans le tableau 4. Ceux-ci doivent être analysés de façon autonome et contrastive. Notons d'abord que les résultats des stratégies exogènes 1 et 2 justifient l'intérêt d'acquérir les informations de sous-catégorisation en 2 étapes (amorçage et consolidation, section 3). Le corpus médical (REA), qui est le plus spécialisé des 4 corpus de test, présente un comportement particulier. Sur ce corpus, les performances des différentes stratégies sont globalement moins bonnes que sur les 3 autres corpus, ce qui illustre le point que nous avons évoqué au début de cet article, à propos de la sensibilité des résultats des analyseurs aux genres des textes. Par ailleurs, la stratégie de base donne de très mauvais résultats sur ce corpus, alors qu'ils sont particulièrement bons sur le corpus littéraire. C'est uniquement sur le corpus médical qu'apparaît, de façon nette, la

nécessité d'exploiter des probabilités de sous-catégorisation spécifiques au corpus (apprentissage endogène). Sur ce corpus, la stratégie endogène donne de meilleurs résultats que la stratégie exogène, et la stratégie mixte est très légèrement supérieure à la stratégie endogène. Sur les corpus littéraire et journaliste, la stratégie exogène est meilleure que la stratégie mixte.

Les ressources de sous-catégorisation syntaxique construites à partir du corpus LM10 sont exploitées par l'analyseur sans avoir été validées manuellement, et les résultats montrent qu'elles sont performantes pour cette tâche. Il convient de préciser que, sur le plan linguistique, ces propriétés de sous-catégorisation ne sont pas comparables aux descriptions que l'on peut trouver dans des lexiques construits à la main, comme le Lexique Grammaire, dans les dictionnaires de langue ou dans les études de psycholinguistique. C'est particulièrement vrai pour les verbes. La probabilité qu'a un verbe de sous-catégoriser telle préposition est calculée à partir de toutes les occurrences (lemmatisées) de ce verbe, sans distinction des différentes acceptions du verbe, alors que l'on sait qu'un même verbe peut avoir des cadres de sous-catégorisation différents selon ses différents sens. Dans le contexte du développement d'un analyseur syntaxique « tout terrain », l'approximation à laquelle conduit ce lissage des sens est un mal nécessaire.

Références

AÏT-MOKTAR S., CHANOD J.-P, ROUX C. (2002), Robustness beyond shallowness : incremental deep parsing, *Natural Language Engineering Journal*, 8(2/3):121-147

BASILI R., PAZIENZA M.-T., VINDIGNI M. (1999), Adaptative Parsing and Lexical Learning, *Proceedings of VEXTAL'99* , Venise

BASILI R., VINDIGNI M. (1998), Adapting a Subcategorization Lexicon to a Domain, *Proceedings of the ECML98 Workshop TANLPS*, Chemnitz, Germany

BOURIGAULT D. (1993), An endogenous Corpus Based Method for Structural Noun Phrase Disambiguation, In *Proceedings of the 6th Conference of the European Chapter of ACL (EACL),* pp. 81-86, Utrecht, The Netherlands

BOURIGAULT D., FABRE C. (2000), Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaire,* 25, Université Toulouse le Mirail, pp. 131-151

CHARNIAK E. (1997), Statistical Parsing with a Contexte-Free Grammar and Word Statistics. *Proceedings of the AAAI97 Conference*, Browne University, Rhode Island, pp.598-603

FABRE C., FRÉROT C (2002), Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus. Actes de la Conférence TALN, pp. 215-224.

FRÉROT C. (2005), *Etude en corpus variés sur l'intégration de ressources linguistiques générales dans un analyseur syntaxique*, Thèse en sciences du langage de l'Université Toulouse le Mirail

FRÉROT C., BOURIGAULT D., FABRE C. (2003), Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus. Le cas du rattachement verbal à distance de la préposition « de », in *Revue t.a.l.*, 44-3

GALA PAVIA N. (2003), *Un modèle d'analyseur syntaxique robuste basé sur la modularité et la lexicalisation de ses grammaires*, PhD, University of Paris XI, Orsay

GILDEA D. (2001), Corpus Variation and Parser Performance. In Lillian Lee and Donna Harma, editors, in *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pp. 167-202

HINDLE D., ROOTH M. (1993), Structural Ambiguity and Lexical Relations. *Computational Linguistics,* 19(1):103-120

KILGARRIFF A., GREFENSTETTE G. (2003), Introduction to the special issue of Web as Corpus. *Computational Linguistics*, 29:3, pp. 333-338

MANNING C. (1993), Automatic Acquisition of Large Subcategorization Dictionary from Corpora, *Proceedings of the 31st Meeting of the Association for Computational Linguistics*, Columbus

PANTEL P., LIN D. (2000), An unsupervised approach to prepositional phrase attachment using contextually similar words. In K. VijayShanker and Chang-Ning Huang, editors, *Proceedings of the 38th Meeting of the Association for Computational Linguistics*, pp. 101-108, Hong Kong

RATNAPARKHI A., REYNAR J., ROUKOS S. (1994), A Maximum Entropy Model for Prepositional Phrase Attachment. *Proceedings of the ARPA Workshop on Human Language Technology*, Morgan Kaufmann

ROLAND D., JURAFSKY, D. (1998). How Verb Subcategorization Frequencies Are Affected By Corpus Choice. *Proceedings of Coling-ACL,* pp. 1122-1128

SEKINE S. (1997), The domain dependence of parsing. *Proceedings of the Fith Conference on Applied Natural Language Processing,* pp. 96-102

SLOCUM J. (1986), How one might Automatically Identify and Adapt to a Sublanguage: An Initial Exploration, in Grishman R. and Kittredge R., eds., *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Lawrence Erlbaum Associates, Hillsdale, N.J., 1986, pp. 195-210

STETINA J., NAGAO M. (1997), Corpus-based PP attachment ambiguity resolution with a semantic dictionary. In J. Zhou and K. Church editors, *Proceedings of the 5th Workshop on Very Large Corpora,* Beijing and Hongkong.